Causal-HalBench: Uncovering LVLMs Object Hallucinations Through Causal Intervention

Zhe Xu [*1], **Zhicai Wang** [*1], **Junkang Wu**[1], **Jinda Lu** [1], **Xiang Wang** [1 †]

[1]University of Science and Technology of China
{xz2002,wangzhic,jkwu0909,lujd}@mail.ustc.edu.cn, xiangwang1223@gmail.com,

Abstract

Large Vision-Language Models (LVLMs) often suffer from object hallucination, making erroneous judgments about the presence of objects in images. We propose this primarily stems from spurious correlations arising when models strongly associate highly co-occurring objects during training, leading to hallucinated objects influenced by visual context. Current benchmarks mainly focus on hallucination detection but lack a formal characterization and quantitative evaluation of spurious correlations in LVLMs. To address this, we introduce causal analysis into the object recognition scenario of LVLMs, establishing a Structural Causal Model (SCM). Utilizing the language of causality, we formally define spurious correlations arising from co-occurrence bias. To quantify the influence induced by these spurious correlations, we develop Causal-HalBench, a benchmark specifically constructed with counterfactual samples and integrated with comprehensive causal metrics designed to assess model robustness against spurious correlations. Concurrently, we propose an extensible pipeline for the construction of these counterfactual samples, leveraging the capabilities of proprietary LVLMs and Text-to-Image (T2I) models for their generation. Our evaluations on mainstream LVLMs using Causal-HalBench demonstrate these models exhibit susceptibility to spurious correlations, albeit to varying extents.

Code — https://github.com/zhexu-ustc/Causal-HalBench

1 Introduction

Recent advancements in Large Language Models (LLMs) (Chiang et al. 2023; Touvron et al. 2023) have driven the emergence of Large Vision-Language Models (LVLMs) (Liu et al. 2023; Dai et al. 2023; Hurst et al. 2024). By integrating a visual encoder with an LLM framework, these models have significantly extended LLMs' capabilities into the visual domain. Consequently, object recognition, a core task in computer vision, has become a crucial metric for evaluating LVLMs' performance. Typically, assessing object recognition in LVLMs involves querying the model about an object's presence, to which it usually responds with a "yes" or "no" (Li et al. 2023; Ye-Bin et al. 2024). However, many studies report that LVLMs frequently suffer from object hallucination, a phenomenon where the models make erroneous judgments about the presence of objects in an input image (Li et al. 2023; Zhai et al. 2023; Zhou et al. 2023). Prior work has indicated that spurious correlations arising from image datasets as a significant factor impeding object recognition (Choi, Torralba, and Willsky 2012; Neuhaus et al. 2023). Specifically, models learn associations between highly co-occurring objects due to imbalanced object distributions within the data. However, these objects are not causally related. This leads to models making predictions based on these learned prior associations, resulting in incorrect judgments. This phenomenon is known as spurious correlations. For example, as depicted in Figure 1 (right), the model might fail to correctly identify skis under a bear's paw due to the absence of a human. Since LVLMs are trained on image data, these spurious correlations are also a primary cause of object hallucination. Yet, accurately quantifying and evaluating their impact remains a significant challenge.

Several efforts have been made to evaluate and understand object hallucination in LVLMs (Li et al. 2023; Ye-Bin et al. 2024; Wang et al. 2023b; Chen et al. 2024a). We categorize these works based on their data construction strategies into two main types: Only Text-axis Evaluation and Text & Image-axis Evaluation. Only Text-axis Evaluation methods directly construct questions from existing image data, such as POPE (Li et al. 2023), NOPE (Lovenia et al. 2023), and CIEM (Hu et al. 2023). For instance, POPE (Figure 1, top-left corner) uses a discriminative question-answering format: "Is there {object} in this image?" and constructs positive and negative questions via image annotations and various negative sampling strategies. Conversely, Text & Image-axis Evaluation methods simultaneously construct both images and questions, including ROPE (Chen et al. 2024a) and BEAF (Ye-Bin et al. 2024). BEAF (Figure 1, bottom-left corner), for example, uses an image editing model to remove objects and then formulates related questions.

Despite their effectiveness in hallucination detection, these methods lack a formal definition of spurious correlations. For instance, BEAF solely assesses a model's perception of image changes through object removal, lacking further discussion on spurious correlations within the model. To address these gaps, we first formalize spurious correlations in LVLMs using a Structural Causal Model (SCM) (Pearl

[*]These authors contributed equally.
[†]Corresponding author

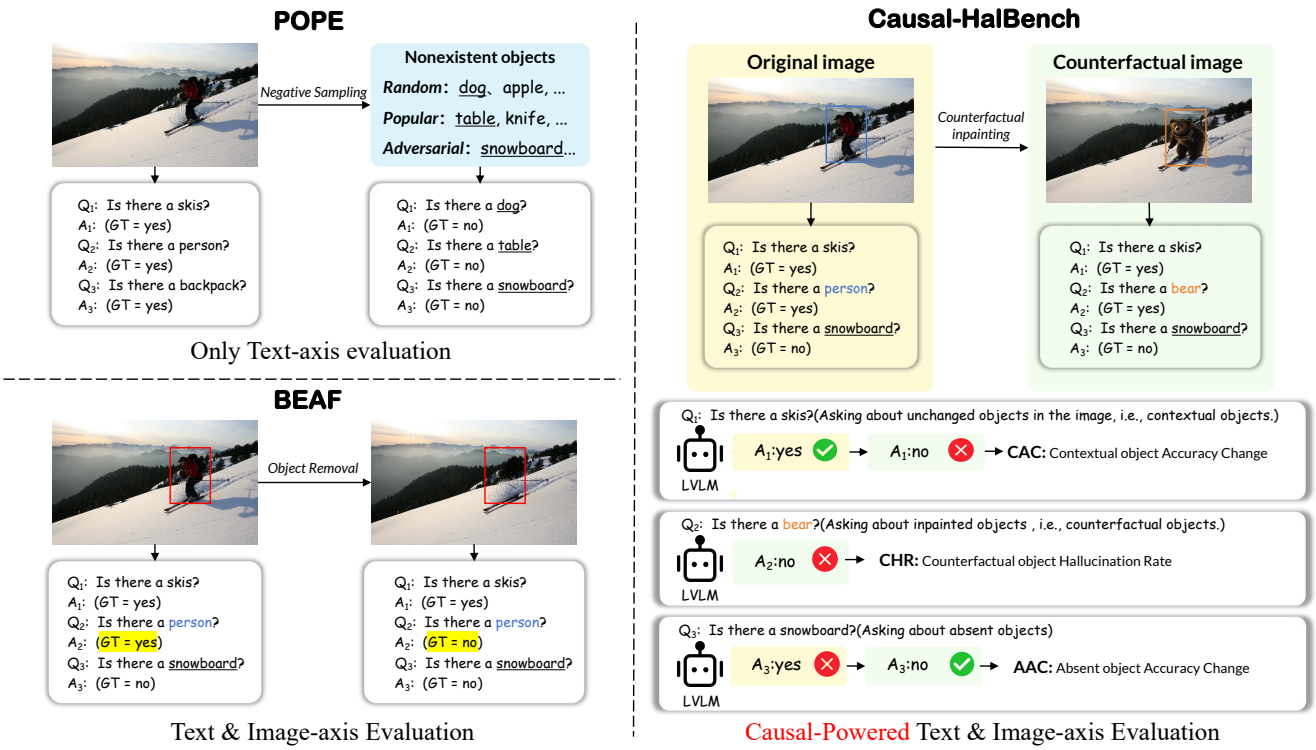


Figure 1: **Illustration of Causal-HalBench.** We present a comparison between current hallucination benchmarks and our Causal-HalBench for Object Hallucination in LVLMs.

2009). Then we propose Causal-HalBench, a causal benchmark for evaluating object hallucination in LVLMs. Inspired by causal intervention (Pearl 2012), we leverage a Visual Content Intervention (VCI) technique to break spurious correlations via introducing counterfactual visual content. To realize VCI, we develop an automated, scalable data synthesis pipeline, which utilizes T2I models to replace high co-occurrence objects with low co-occurrence objects. As illustrated in Figure 1 (right), we construct 1387 counterfactual images from 757 original images via counterfactual inpainting, yielding 9709 corresponding Question-Answering (QA) pairs. We also propose novel causal-based metrics to quantify the impact of spurious correlations by observing model responses to images before and after changes. Our experiments on mainstream LVLMs reveal that spurious correlations are widespread in LVLMs, leading to hallucination across various aspects. Furthermore, we find that newer models may be more susceptible to spurious correlations.

In summary, our main contributions are:

- We are the first to integrate causal analysis into LVLM object hallucination study, systematically analyzing and quantifying spurious correlation influence using SCMs.
- We develop an automated and scalable pipeline leveraging T2I models and proprietary LVLMs for high-quality causal counterfactual sample generation.
- We establish Causal-HalBench, the first causal hallucination detection benchmark, comprising over 10,00 counterfactual samples and comprehensive evaluation metrics. Through extensive experiments on Causal-HalBench, we provide empirical evidence of state-of-the-art LVLMs' susceptibility to spurious correlations, identifying a key area for improving model faithfulness.

2 Related Work

Large Vision-Language Models. The past few years have witnessed rapid advancements in Large Vision-Language Models (LVLMs), which integrate powerful visual encoders with large language model (LLM) architectures to achieve multimodal understanding and generation capabilities. Early pioneers like CLIP (Radford et al. 2021) and ALIGN (Jia et al. 2021) demonstrate robust vision-language alignment. More recent advancements, such as LLaVA (Liu et al. 2023), Mplug-Owl(Ye et al. 2023) and GPT-4o (Hurst et al. 2024) have pushed the boundaries, exhibiting remarkable zero-shot and few-shot performance in various visual tasks, such as visual question answering (Hudson and Manning 2019) and image captioning (Agrawal et al. 2019).

Object Hallucination in LVLMs. Despite impressive capabilities, LVLMs often generate objects inconsistent with images, a problem known as object hallucination (Li et al. 2023; Zhai et al. 2023). Several methods have been suggested to mitigate the object hallucination issue (Zhai et al. 2023; Gunjal, Yin, and Bas 2024). To quantify progress on mitigating them, various benchmarks have been developed (Rohrbach et al. 2018; Li et al. 2023; Ye-Bin et al. 2024; Wang et al. 2023b; Chen et al. 2024a). While effective for

detection, these benchmarks lack a causal analysis framework to quantify the causal effect strength of specific spurious correlations on predictions. This limits a deeper, causal understanding of hallucination, which our work addresses.

Spurious Correlations/Biases in Language Model. Spurious correlations constitute a critical issue in LVLMs: models often learn misleading statistical associations between co-occurring objects in training data (without true causal relationships), impairing their correct prediction ability and significantly contributing to problems such as object hallucination. While causal analysis has emerged as a powerful tool for addressing analogous biases in Natural Language Processing (NLP) (Wang et al. 2022; Zhu et al. 2022; Wang et al. 2023a,c), this approach remains underutilized for spurious correlation-induced hallucination in LVLM. Although benchmarks like MM-SPUBENCH (Ye et al. 2024b) and other existing methods (Li et al. 2023; Ye-Bin et al. 2024; Wang et al. 2023b) explore spurious biases, they primarily focus on detecting bias, yet lack the capacity for rigorous causal analysis and examination of these spurious biases(Zhu et al. 2024a,b). Our work directly addresses this critical gap by explicitly applying a causal framework to examine and quantify the causal impact of spurious correlations on object hallucination in LVLMs.

3 Methodology

In this section, we first leverage SCM to formalize spurious correlations of object hallucination in LVLMs (§3.1). Then we introduce Visual Content Intervention (VCI), a curated causal intervention technique that breaks the non-causal path by manipulating the visual content of images. This enables us to quantify their causal effect using Average Causal Effect (ACE) and Direct Causal Effect (DCS) (§3.2).

3.1 Causal Analysis of Object Hallucination

To analyze how spurious correlations cause object hallucination in LVLMs, we propose a SCM, a powerful tool that leverages structural theory to estimate causal effects from data (Pearl et al. 2000), as shown in Figure 2 (left). In this SCM, X represents the input image, Q denotes the question provided to the LVLM (e.g., "Is there {object} in the image?"), and Y signifies the LVLM's output regarding object existence. A key variable is C, referred to as **Co-occurrence Bias**. This bias denotes the statistical patterns (from training data) of frequent co-occurrence of specific objects/concepts, reflecting the model's tendency to associate these typically non-causal learned relationships. Ideally, an LVLM should predict accurately based only on X (the direct causal path $X \rightarrow Y$), with Q directly influencing Y via $Q \rightarrow Y$.

However, as Figure 2 (left) illustrates, C functions as a confounder in the relationship between X and Y (Pearl 2009)-a variable that influences both the input and the output, causing a spurious correlation between them. In our SCM, the link $C \rightarrow X$ shows that input images (X), particularly those in the training data, inherently contain co-occurrence patterns driven by this Co-occurrence Bias (C); while the path $C \rightarrow Y$ indicates that the model directly uses this Co-occurrence Bias (C) to generate its output (Y), serving as a cognitive shortcut that bypasses in-depth analysis of X. This confounding effect creates a backdoor path: $X \leftarrow C \rightarrow Y$. Spurious correlations form through this path. When an LVLM over-relies on this pathway, it may predominantly leverage Co-occurrence Bias (C) embedded in X rather than the visual information within X to predict the presence of an object (Y), thereby leading to hallucinations.

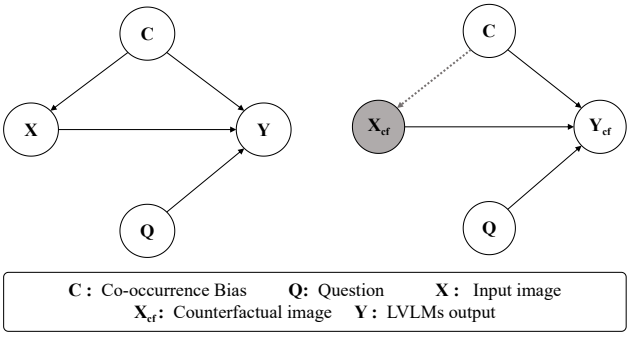


Figure 2: **Illustration of SCM.** We present the original SCM of spurious correlation (left), contrasted with the SCM derived via VCI (right).

3.2 Visual Content Intervention

In the causal analysis presented in the previous section, we treat C as a confounder between X and Y. Prior research has extensively explored confounder analysis (Keith, Jensen, and O'Connor 2020; Qian et al. 2021; Feder et al. 2022; Weld et al. 2022), and central to these methods is the concept of intervention (denoted as $\text{do}(\cdot)$) (Pearl 2012). Intervention refers to the manipulation of a specific node (variable) within a causal system, which effectively wipes out all its incoming causal links and sets it to a specific value, thereby allowing us to isolate its true, unconfounded effects.

We propose **Visual Content Intervention (VCI)** as our tailored method for spurious correlations. The causal mechanism of VCI is depicted in Figure 2 (right). Specifically, VCI systematically constructs a counterfactual image (x_{cf}) from an original image by introducing counterfactual visual content. This strategic modification of X disrupts the influence of C inherent in the original image, by altering its associated visual cues. The formal expression for the outcome under this intervention is:

$$y_{cf} = Y(\text{do}(X = x_{cf})). \tag{1}$$

The comparison between the intervened and original outcomes , central to counterfactual analysis (Pearl 2019), isolates the causal effect of spurious correlations. The average of these isolated effects defines the **Average Causal Effect (ACE)** (Rubin 1974), a fundamental causal inference concept, We formally define ACE as follows:

$$ACE = E[Y \mid \text{do}(X = X_{cf}), Q] - E[Y \mid X, Q]. \tag{2}$$

Beyond ACE, we also introduce the **Direct Causal Strength (DCS)**. DCS quantifies the model's reliance on actual visual information under controlled spurious correlations. It is formally represented by the expected outcome under intervention:

$$DCS = E[Y \mid \text{do}(X = x_{cf}), Q]. \tag{3}$$

ACE and DCS form the fundamental causal basis for quantifying the effects of spurious correlations. By introducing specific definitions, we operationalize these concepts into measurable metrics for our benchmark. These will be detailed in the subsequent section.

4 Causal-HalBench

Based on the VCI proposed in the preceding section, we utilize inpainting models and LVLMs to develop an automated, scalable pipeline for high-quality counterfactual sample generation (§4.1). Subsequently, based on our proposed pipeline, we establish Causal-HalBench and provide an overview of our dataset (§4.2). Finally, we construct comprehensive metrics based on our prior causal analysis to quantify the effect of spurious correlation in LVLMs (§4.3).

4.1 Dataset Construction Pipeline

To implement Visual Content Intervention (VCI) and generate counterfactual samples (X_{cf}) for causal analysis, we design an automated and scalable pipeline. This three-stage pipeline, illustrated in Figure 3, systematically creates counterfactual samples through: 1) Intervention Objects Selection, 2) Counterfactual Description Generation, and 3) Counterfactual Inpainting. We offer a detailed explanation of each stage as follows. For more details, please refer to the Appendix B.1.

Intervention Objects Selection. The initial stage of our pipeline is critical for identifying intervention objects that fulfill two key requirements for effective counterfactual sample generation: ensuring low co-occurrence with the original context and enhancing inpainting success. This process begins by designating each annotated object in the original image as a target object o_t. For every o_t, we employ **Contextual Sampling** to probabilistically sample a contextual object o_c based on its co-occurrence frequency with o_t , ensuring visual relevance and diversity. Subsequently, for **Counterfactual Selection**, we form a candidate pool of top-k objects with the lowest co-occurrence frequencies with o_c. Given the original image with the highlighted target object, the powerful proprietary LVLM Gemini (Team et al. 2023) then selects the most appropriate object from this candidate pool to replace o_t, prioritizing visual suitability to ensure the resulting counterfactual object o_{cf} is not only a low-association adversarial visual element but also highly amenable to inpainting, significantly increasing the inpainting success rate. This meticulous approach ultimately yields (o_t, o_c, o_{cf}) triplets, forming the indispensable basis for subsequent counterfactual sample generation.

Counterfactual Description Generation. The second stage of our pipeline aims to create accurate, high-quality counterfactual descriptions that will serve as inpainting prompts.

	Image	Img-Q Pair	Yes	No
Original	757	4161	1387	2774
Counterfactual	1387	5548	2774	2774
Total	2144	9709	4161	5548

Table 1: **Dataset Statistics.** Causal-HalBench contains 9.7K image-question pairs, consisting of the original and counterfactual ones.

These descriptions are crafted with the help of Gemini. Specifically, Gemini is first prompted to generate a description of the original image based on its ground-truth object annotations. Subsequently, leveraging this initial description, Gemini is further prompted to incorporate the conceptual replacement of the target object o_t with the previously selected counterfactual object o_{cf}, thereby constructing a reasonable and accurate counterfactual description. This ensures the generated description accurately reflects the desired visual modification for the subsequent inpainting process.

Counterfactual Inpainting. The final stage of our pipeline is dedicated to generating high-quality counterfactual samples (X_{cf}). For this purpose, we employ FLUX-controlnet inpainting (Zhang, Rao, and Agrawala 2023; Labs 2024) as our core inpainting model. We first utilize the Segment Anything Model (SAM) (Kirillov et al. 2023) to extract precise masks for the objects. Then, to ensure comprehensive coverage and minimize the potential influence of specific object shapes, these masks are subsequently dilated. Finally, the original image, the dilated mask, and the counterfactual description are input into the inpainting model, yielding the desired counterfactual image.

Discussion. We acknowledge that using an inpainting model is an approximation of the theoretical $do(X = x_{cf})$ intervention. Potential visual artifacts introduced by the inpainting process could act as interfering factors, such as unnatural foreground-background integration and the disruption of context objects. To mitigate this, we employed several strategies: (1) using a state-of-the-art inpainting model (FLUX-controlnet) to ensure high fidelity; (2) performing manual filtering to discard low-quality images; and (3) as we will show in Section 5.3, we use CLIP Score to quantitatively verify that the synthetic region correctly represents the counterfactual object while diminishing the signal of the original target object (Ye-Bin et al. 2024), thus providing evidence for the effectiveness of our intervention approximation.

4.2 Dataset Overview

Our dataset is made up of images and their corresponding Q&A pairs. The images are split into original and counterfactual categories. We first randomly sample original images from the MSCOCO dataset's validation set (Lin et al. 2014) and then use our designed pipeline to generate corresponding counterfactual images. After a round of manual filtering, we finally obtain a total of 757 original images and 1387 generated counterfactual images. We adopt the question format "Is there {object} in this image?", consistent

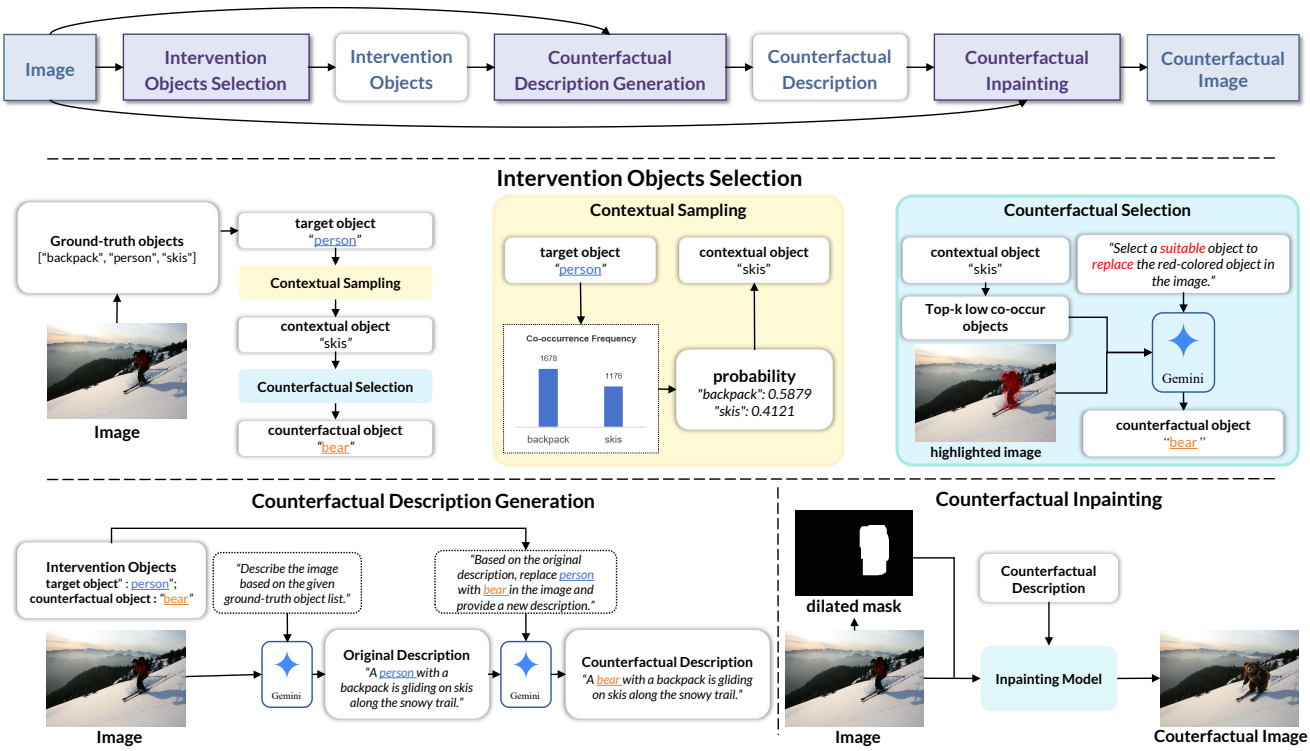


Figure 3: **The Data Construction Pipeline of Causal-HalBench.** We propose a three-stage, fully automated, and scalable pipeline for constructing counterfactual samples in Causal-HalBench.

with POPE(Li et al. 2023). For positive questions where the answer is "yes", we query the contextual object o_c for both original and counterfactual images, and the counterfactual object o_{cf} specifically for counterfactual images. For negative questions where the answer is "no", we query the top-k absent objects (o_a) that frequently appear with the original target object (o_t). Ultimately, our dataset includes 2144 images and 9709 image-question pairs; see Table 1 for specific statistical information.

4.3 Evaluation Metrics

Building upon the concepts of ACE and DCS, we design our evaluation metrics. We first establish three question types: **contextual questions** (Q_c) for a contextual object o_c (GT: Yes); **counterfactual questions** (Q_{cf}) for a counterfactual object o_{cf} (GT: Yes); and **absent questions** (Q_a) for an object o_a (GT: No). We then formalize the model's output, Y, as a binary variable, where Y=1 for a correct answer and Y=0 for an incorrect one. This allows the expectation, $E[Y]$, within the ACE and DCS to be directly interpreted as the model's accuracy. We define the model's overall accuracy on an image set X and a question set Q as $Acc(X, Q)$. Based on these definitions, we formulate our metrics as follows.

Contextual object Accuracy Change (CAC). CAC quantifies the decrease in accuracy of Q_c from the original image to the counterfactual image, which represents ACE for $Q = Q_c$.

$$CAC = Acc(X, Q_c) - Acc(X_{cf}, Q_c). \tag{4}$$

A higher CAC reflects a stronger causal effect of spurious correlations, indicating that the model is more prone to making predictions based on co-occurrence biases.

Absent object Accuracy Change (AAC). AAC quantifies the increase in accuracy of Q_a from the original image to the counterfactual image, which represents ACE for $Q = Q_a$.

$$AAC = Acc(X_{cf}, Q_a) - Acc(X, Q_a). \tag{5}$$

A higher AAC indicates that the model is more influenced by spurious correlations, leading it to hallucinate related nonexistent objects.

Counterfactual object Hallucination Rate (CHR). CHR measures the hallucination rate for the counterfactual object on counterfactual images, which reflects DCS, with $Acc(X_{cf}, Q_{cf})$ corresponding to DCS for $Q = Q_{cf}$.

$$CHR = 1 - Acc(X_{cf}, Q_{cf}). \tag{6}$$

A lower CHR reflects the model's strong visual perception and weaker influence from spurious correlations.

5 Experiment

In this section, we conduct a series of experiments on mainstream LVLMs leveraging Causal-HalBench. First, we evaluate various LVLMs on Causal-HalBench using our proposed causality-based metrics and analyze the experimental results (§5.1). Subsequently, we compare our Causal-HalBench with representative object hallucination benchmarks, encompassing both discriminative (POPE (Li et al.

Model	Original Image		Counterfactual Image		Metrics			
	Acc(Q_c)↑	Acc(Q_a)↑	Acc(Q_c)↑	Acc(Q_a)↑	Acc↑	ΔAcc(Q_c)↓	ΔAcc(Q_a)↓	CHR↓
LLaVA-NEXT-8B	90.9	81.1	86.4	82.2	85.9	4.5	1.1	6.8
LLaVA-onevision-7B	82.8	94.8	79.2	95.0	87.0	3.6	0.2	14.4
Kimi-VL-A3B	91.0	75.2	85.6	85.4	84.4	5.4	10.2	7.4
MiniCPM-o-2_6	89.8	78.1	86.6	80.9	83.1	3.2	2.8	14.9
InternVL2.5-8B	88.9	72.4	86.1	79.5	78.8	2.8	7.1	29.3
mPLUG-Owl3-7B	81.8	93.7	77.9	94.0	86.6	3.8	0.3	14.4
Qwen2.5-VL-7B	70.0	97.6	68.2	98.1	84.1	1.8	0.5	27.3
GPT-4o	79.1	91.0	75.5	92.5	84.3	3.6	1.5	12.4
Gemini1.5-pro	81.9	87.4	73.8	87.7	86.1	8.1	0.3	21.4

Table 2: **Evaluation Results on Causal-HalBench.** We evaluate various LVLMs on Causal-HalBench. Acc(Q_c) and Acc(Q_a) report the model's accuracy on contextual objects and absent objects in the original and counterfactual images, while ΔAcc(Q_c) and ΔAcc(Q_a) indicate the resulting changes. All values in the table are percentages. Underlined values indicate the best results.

2023)) and generative (CHAIR (Rohrbach et al. 2018)) approaches (§5.2). Finally, we perform additional study on Causal-HalBench, visualizing its co-occurrence patterns and assessing the dataset quality utilizing CLIP Score (Hessel et al. 2021) (§5.3).

Implementation Details. All the reported performance is the zero-shot performance of LVLMs on a single A40, including GPT-4o (Hurst et al. 2024), Gemini1.5-pro (Team et al. 2024), LLAvA-NEXT (Liu et al. 2024), LLAvA-onevision (Li et al. 2024), Kimi-VL (Team et al. 2025), MiniCPM (Team 2025), InternVL2.5 (Chen et al. 2024b), mPLUG-Owl3 (Ye et al. 2024a), and Qwen2.5-VL (Bai et al. 2025). For more details, please refer to the Appendix B.2.

5.1 Evaluation Results on Causal-HalBench

We evaluate several LVLMs on Causal-HalBench and analyze the results. Experimental results are shown in Table 2.

Main Results. As presented in Table 2, the majority of models demonstrate high CAC values. Notably, Qwen2.5-VL-7B achieve the lowest CAC at 1.8%, yet this desirable outcome is accompanied by a diminished Acc(Q_c) compared to other models. Conversely, in the evaluation of AAC, Kimi-VL-A3B record the highest AAC value at 10.2%, substantially exceeding all other models; InternVL2.5-8B (7.1%) and MiniCPM-o-2_6 (2.8%) followed, while the remaining models demonstrate lower AAC values. Finally, regarding CHR, LLaVA-NEXT-8B exhibit the optimal performance at 6.8%, with Kimi-VL-A3B also demonstrating commendable results at 7.4%. In contrast, InternVL2.5-8B (29.3%) and Qwen2.5-VL-7B (27.3%) perform notably worse in this regard. Meanwhile, the overall performance of closed-source models do not significantly outperform open-source models.

Analysis. Our experimental results indicate that LVLMs are susceptible to spurious correlations to varying extents, with nearly all models exhibiting high CAC values, confirming their general vulnerability to spurious correlations when contextual elements are altered. We also observe that the influence of spurious correlations on models is multi-faceted. For instance, while Kimi-VL-A3B demonstrates exceptional performance in CHR, it exhibits the poorest performance in AAC. Conversely, Qwen2.5-VL-7B performs favorably on AAC but shows noticeably inferior results in CHR . This underscores the inherent limitations of relying on a single accuracy metric in traditional evaluation methodologies, thereby necessitating a comprehensive evaluation across multiple metrics to thoroughly assess spurious correlations within these models. Furthermore, the aforementioned findings suggest that while more recent models (e.g., Qwen2.5-VL) may outperform earlier iterations (e.g., LLaVA-NEXT) on certain general benchmarks, their susceptibility to spurious correlations can be markedly more pronounced. This phenomenon can be attributed to the increased scale of training data, thereby underscoring the critical importance of detecting and mitigating spurious correlations for the continued advancement of LVLMs.

5.2 Comparison with Other Benchmarks

To validate the efficacy of our benchmark, we conduct a comparative analysis of Causal-HalBench against representative object hallucination benchmarks, specifically including both discriminative (POPE(Li et al. 2023)) and generative (CHAIR(Rohrbach et al. 2018))) approaches.

Comparison with POPE. In the Table 3, we evaluate model performance on both POPE and the entire Causal-HalBench, using accuracy as the primary evaluation metric. A discernible consistency in performance trends emerge between the two benchmarks. Critically, however, the deliberate inclusion of counterfactual samples in Causal-HalBench leads to a subdued performance compared to POPE. This finding affirms Causal-HalBench's robustness for conventional evaluation, while simultaneously underscoring its superior extensibility beyond POPE, particularly through its integrated, causality-based metrics designed for a precise quantification of spurious correlations.

Comparison with CHAIR. Beyond discriminative testing, we also conduct open generation testing on Causal-HalBench. For each image, we prompt the model with "Provide a brief description of the given image." In Tables 4, we present the CHAIR results for both the full dataset and a subset containing only counterfactual images, alongside the CHR results. CHAIR is a popular metric for evaluating

Model	Causal-HalBench	POPE
LLaVA-NEXT-8B	85.9	87.1
LLaVA-onevision-7B	87.0	88.4
Kimi-VL-A3B	84.4	88.5
MiniCPM-o-2_6	83.1	86.7
InternVL2.5-8B	78.8	88.7
mPLUG-Owl3-7B	86.6	87.1
Qwen2.5-VL-7B	84.1	85.9

Table 3: **Comparison of Performance on POPE and Causal-HalBench.** We compare model accuracy on POPE and the entire Causal-HalBench.

Model	CHAIR-S↓	CHAIR-I↓	CHR↓
Qwen2.5-VL-7B	44.1 (54.7)	15.8 (20.0)	27.3
InternVL2.5-8B	68.2 (76.2)	18.7 (22.7)	29.3
MiniCPM-o-2_6	22.9 (30.0)	13.7 (17.8)	14.9
LLaVA-onevision-7B	19.3 (26.9)	11.1 (15.2)	14.4
mPLUG-Owl3-7B	24.5 (33.6)	14.3 (18.4)	14.4
Kimi-VL-A3B	34.9 (43.1)	14.6 (18.1)	7.4
LLaVA-NEXT-8B	29.0 (38.8)	13.7 (18.1)	6.8

Table 4: **Comparison of CHAIR metrics and CHR on Causal-HalBench.** We measure the CHAIR score based on the answers derived from the images in BEAF using the prompt "Provide a brief description of the given image." Note: Results in parentheses are measured on a subset containing only counterfactual images within Causal-HalBench.

object hallucination in image captioning tasks. It comprises CHAIR-I and CHAIR-S, which respectively count the hallucinated instances and sentences containing these objects in the generated output. The lower the score, the better. From the results, we observe that if the model includes more hallucinated objects and sentences in its generated output, then CHR tends to be higher, which demonstrates the effectiveness of our evaluation method. Simultaneously, the model's hallucination problem worsened during open generation for counterfactual images, further illustrating the effect of spurious correlations on model hallucinations.

5.3 Additional Study on Causal-HalBench

Balanced Co-occurrence Patterns. To systematically evaluate the effects of spurious correlations within LVLMs, our proposed Causal-HalBench intentionally enriches and balances co-occurrence patterns in image data through VCI. To visualize this comparison, we calculate the co-occurrence matrices between objects in both the original and modified image datasets and plot their respective heatmaps, as shown in Figure 4. We clearly observe a distinct shift: compared to the original image dataset, which exhibits significantly concentrated high-frequency co-occurrence patterns between certain objects (manifesting as darker regions), the modified image dataset demonstrably increases the frequency of previously low-frequency or even non-co-occurring object pairs (indicated by a greater variety of shaded regions rather than uniform white). Concurrently, this process effectively dilutes the strength of pre-existing strong correlations, ultimately presenting a more balanced distribution of co-occurrence patterns.

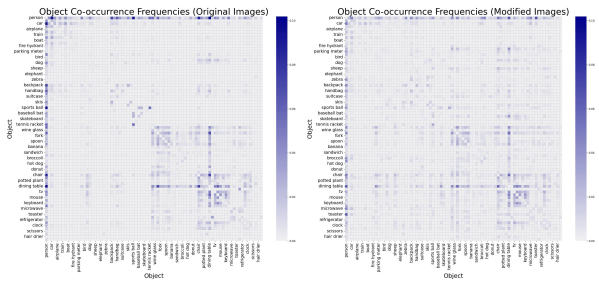


Figure 4: **Visualization of Co-occurrence Patterns.** We present the heatmap of the co-occurrence matrix for original images (left), contrasted with the heatmap of the co-occurrence matrix for modified images (right).

Synthetic Data Quality Evaluation. We manually filter the dataset to ensure the quality of the synthesized data. However, considering that most current LVLMs adopt CLIP as their visual encoder, we further evaluate the quality of the synthesized data using CLIP Score. We measure the CLIP score between an object and a prompt "a photo of {object}". We crop the inpainting region of the image and set the {object} in the prompt to be the target object and the counterfactual object, respectively. As shown in Table 5, the synthesized images have a lower CLIP Score for the target object and a higher CLIP Score for the counterfactual object compared to the original images, which ensures the quality of the synthesized data.

Type	Target	Counterfactual
Original	26.5	20.6
Synthetic	22.4	27.5

Table 5: **CLIP Score Results of Causal-HalBench.** We measure the CLIP Score of original and counterfactual images in Causal-HalBench with the text prompt "a photo of {object}". Here, object includes both the target object and the counterfactual object.

6 Conclusion

Our study first introduces causal analysis to LVLM hallucination research. We propose that during training, models learn co-occurrence bias between highly co-occurring objects, leading to misjudgments in object recognition. This phenomenon is termed spurious correlation. To evaluate this, we create a new benchmark called Causal-HalBench. It uses advanced proprietary LVLMs and inpainting models to construct counterfactual samples, thereby breaking spurious correlations in images, and introduces new causality-based metrics to quantify the impact of these spurious correlations. Experiments show that all current models are affected by this issue to various degree, highlighting a key way to make them more accurate in the future.

Acknowledgments

This research is supported by the National Science and Technology Major Project (2023ZD0121102).

References

Agrawal, H.; Desai, K.; Wang, Y.; Chen, X.; Jain, R.; Johnson, M.; Batra, D.; Parikh, D.; Lee, S.; and Anderson, P. 2019. Nocaps: Novel object captioning at scale. In *Proceedings of the IEEE/CVF international conference on computer vision*, 8948–8957.

Bai, S.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Song, S.; Dang, K.; Wang, P.; Wang, S.; Tang, J.; et al. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Chen, X.; Ma, Z.; Zhang, X.; Xu, S.; Qian, S.; Yang, J.; Fouhey, D.; and Chai, J. 2024a. Multi-object hallucination in vision language models. *Advances in Neural Information Processing Systems*, 37: 44393–44418.

Chen, Z.; Wang, W.; Cao, Y.; Liu, Y.; Gao, Z.; Cui, E.; Zhu, J.; Ye, S.; Tian, H.; Liu, Z.; et al. 2024b. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*.

Chiang, W.-L.; Li, Z.; Lin, Z.; Sheng, Y.; Wu, Z.; Zhang, H.; Zheng, L.; Zhuang, S.; Zhuang, Y.; Gonzalez, J. E.; et al. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*, 2(3): 6.

Choi, M. J.; Torralba, A.; and Willsky, A. S. 2012. Context models and out-of-context objects. *Pattern Recognition Letters*, 33(7): 853–862.

Dai, W.; Li, J.; Li, D.; Tiong, A.; Zhao, J.; Wang, W.; Li, B.; Fung, P. N.; and Hoi, S. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in neural information processing systems*, 36: 49250–49267.

Feder, A.; Keith, K. A.; Manzoor, E.; Pryzant, R.; Sridhar, D.; Wood-Doughty, Z.; Eisenstein, J.; Grimmer, J.; Reichart, R.; Roberts, M. E.; et al. 2022. Causal inference in natural language processing: Estimation, prediction, interpretation and beyond. *Transactions of the Association for Computational Linguistics*, 10: 1138–1158.

Gunjal, A.; Yin, J.; and Bas, E. 2024. Detecting and preventing hallucinations in large vision language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 18135–18143.

Hessel, J.; Holtzman, A.; Forbes, M.; Bras, R. L.; and Choi, Y. 2021. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*.

Hu, H.; Zhang, J.; Zhao, M.; and Sun, Z. 2023. Ciem: Contrastive instruction evaluation method for better instruction tuning. *arXiv preprint arXiv:2309.02301*.

Hudson, D. A.; and Manning, C. D. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 6700–6709.

Hurst, A.; Lerer, A.; Goucher, A. P.; Perelman, A.; Ramesh, A.; Clark, A.; Ostrow, A.; Welihinda, A.; Hayes, A.; Radford, A.; et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Jia, C.; Yang, Y.; Xia, Y.; Chen, Y.-T.; Parekh, Z.; Pham, H.; Le, Q.; Sung, Y.-H.; Li, Z.; and Duerig, T. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, 4904–4916. PMLR.

Keith, K. A.; Jensen, D.; and O'Connor, B. 2020. Text and causal inference: A review of using text to remove confounding from causal estimates. *arXiv preprint arXiv:2005.00649*.

Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.-Y.; et al. 2023. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, 4015–4026.

Labs, B. F. 2024. FLUX. https://github.com/black-forest-labs/flux.

Li, B.; Zhang, Y.; Guo, D.; Zhang, R.; Li, F.; Zhang, H.; Zhang, K.; Zhang, P.; Li, Y.; Liu, Z.; et al. 2024. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*.

Li, Y.; Du, Y.; Zhou, K.; Wang, J.; Zhao, W. X.; and Wen, J.-R. 2023. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, 740–755. Springer.

Liu, H.; Li, C.; Li, Y.; Li, B.; Zhang, Y.; Shen, S.; and Lee, Y. J. 2024. Llavanext: Improved reasoning, ocr, and world knowledge.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023. Visual instruction tuning. *Advances in neural information processing systems*, 36: 34892–34916.

Lovenia, H.; Dai, W.; Cahyawijaya, S.; Ji, Z.; and Fung, P. 2023. Negative object presence evaluation (nope) to measure object hallucination in vision-language models. *arXiv preprint arXiv:2310.05338*.

Neuhaus, Y.; Augustin, M.; Boreiko, V.; and Hein, M. 2023. Spurious features everywhere-large-scale detection of harmful spurious features in imagenet. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 20235–20246.

Pearl, J. 2009. Causal inference in statistics: An overview.

Pearl, J. 2012. The do-calculus revisited. *arXiv preprint arXiv:1210.4852*.

Pearl, J. 2019. Causal and counterfactual inference. The handbook of rationality.

Pearl, J.; et al. 2000. Models, reasoning and inference. *Cambridge, UK: CambridgeUniversityPress*, 19(2).

Qian, C.; Feng, F.; Wen, L.; Ma, C.; and Xie, P. 2021. Counterfactual inference for text classification debiasing. In *Proceedings of the 59th Annual Meeting of the Association for*

Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), 5434–5445.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PmLR.

Rohrbach, A.; Hendricks, L. A.; Burns, K.; Darrell, T.; and Saenko, K. 2018. Object hallucination in image captioning. *arXiv preprint arXiv:1809.02156*.

Rubin, D. B. 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5): 688.

Team, G.; Anil, R.; Borgeaud, S.; Alayrac, J.-B.; Yu, J.; Soricut, R.; Schalkwyk, J.; Dai, A. M.; Hauth, A.; Millican, K.; et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Team, G.; Georgiev, P.; Lei, V. I.; Burnell, R.; Bai, L.; Gulati, A.; Tanzer, G.; Vincent, D.; Pan, Z.; Wang, S.; et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Team, K.; Du, A.; Yin, B.; Xing, B.; Qu, B.; Wang, B.; Chen, C.; Zhang, C.; Du, C.; Wei, C.; et al. 2025. Kimi-vl technical report. *arXiv preprint arXiv:2504.07491*.

Team, O. M.-o. 2025. Minicpm-o 2.6: A gpt-4o level mllm for vision, speech, and multimodal live streaming on your phone.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Wang, F.; Mo, W.; Wang, Y.; Zhou, W.; and Chen, M. 2023a. A causal view of entity bias in (large) language models. *arXiv preprint arXiv:2305.14695*.

Wang, J.; Wang, Y.; Xu, G.; Zhang, J.; Gu, Y.; Jia, H.; Wang, J.; Xu, H.; Yan, M.; Zhang, J.; et al. 2023b. Amber: An llm-free multi-dimensional benchmark for mllms hallucination evaluation. *arXiv preprint arXiv:2311.07397*.

Wang, Y.; Chen, M.; Zhou, W.; Cai, Y.; Liang, Y.; Liu, D.; Yang, B.; Liu, J.; and Hooi, B. 2022. Should we rely on entity mentions for relation extraction? debiasing relation extraction with counterfactual analysis. *arXiv preprint arXiv:2205.03784*.

Wang, Y.; Hooi, B.; Wang, F.; Cai, Y.; Liang, Y.; Zhou, W.; Tang, J.; Duan, M.; and Chen, M. 2023c. How Fragile is Relation Extraction under Entity Replacements? *arXiv preprint arXiv:2305.13551*.

Weld, G.; West, P.; Glenski, M.; Arbour, D.; Rossi, R. A.; and Althoff, T. 2022. Adjusting for confounders with text: Challenges and an empirical evaluation framework for causal inference. In *Proceedings of the international AAAI conference on web and social media*, volume 16, 1109–1120.

Ye, J.; Xu, H.; Liu, H.; Hu, A.; Yan, M.; Qian, Q.; Zhang, J.; Huang, F.; and Zhou, J. 2024a. mplug-owl3: Towards long image-sequence understanding in multi-modal large language models. *arXiv preprint arXiv:2408.04840*.

Ye, Q.; Xu, H.; Xu, G.; Ye, J.; Yan, M.; Zhou, Y.; Wang, J.; Hu, A.; Shi, P.; Shi, Y.; et al. 2023. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*.

Ye, W.; Zheng, G.; Ma, Y.; Cao, X.; Lai, B.; Rehg, J. M.; and Zhang, A. 2024b. Mm-spubench: Towards better understanding of spurious biases in multimodal llms. *arXiv preprint arXiv:2406.17126*.

Ye-Bin, M.; Hyeon-Woo, N.; Choi, W.; and Oh, T.-H. 2024. Beaf: Observing before-after changes to evaluate hallucination in vision-language models. In *European Conference on Computer Vision*, 232–248. Springer.

Zhai, B.; Yang, S.; Zhao, X.; Xu, C.; Shen, S.; Zhao, D.; Keutzer, K.; Li, M.; Yan, T.; and Fan, X. 2023. HallE-Switch: Rethinking and Controlling Object Existence Hallucinations in Large Vision-Language Models for Detailed Caption.

Zhang, L.; Rao, A.; and Agrawala, M. 2023. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, 3836–3847.

Zhou, Y.; Cui, C.; Yoon, J.; Zhang, L.; Deng, Z.; Finn, C.; Bansal, M.; and Yao, H. 2023. Analyzing and mitigating object hallucination in large vision-language models. *arXiv preprint arXiv:2310.00754*.

Zhu, X.; Wang, S.; Lu, J.; Hao, Y.; Liu, H.; and He, X. 2024a. Boosting few-shot learning via attentive feature regularization. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, 7793–7801.

Zhu, X.; Zhu, B.; Tan, Y.; Wang, S.; Hao, Y.; and Zhang, H. 2024b. Enhancing zero-shot vision models by label-free prompt distribution learning and bias correcting. *Advances in Neural Information Processing Systems*, 37: 2001–2025.

Zhu, Y.; Sheng, Q.; Cao, J.; Li, S.; Wang, D.; and Zhuang, F. 2022. Generalizing to the future: Mitigating entity bias in fake news detection. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2120–2125.

Appendix

Contents

A Discussions

A.1 Limitations.

While our work introduces an automated and scalable pipeline for counterfactual sample generation, which incorporates diverse strategies to enhance its efficacy, the current success rate is still limited to around 40%. This is primarily due to constraints from MSCOCO [6] dataset annotation errors and limitations in the generative models' capabilities. Notably, the success rate plummets when attempting to replace smaller objects, thereby imposing a significant burden of manual filtering. Although such limitations exist, these aspects are expected to naturally diminish or disappear with the advancement of in-painting models and the optimization of image annotation. We plan to optimize the pipeline's robustness in our future work.

A.2 Ethic Considerations

The MSCOCO dataset that we used to curate Causal-HalBench adhere to strict guidelines to exclude any harmful, unethical, or offensive content. Furthermore, during our manual filtering process, human reviewers are instructed to remove any personally identifiable or offensive content. Finally, we do not generate any harmful, ethical, or offensive content in our dataset.

A.3 Artifacts and licenses

We report a list of licenses for all datasets and models used in our experiment in Table 1. We strictly follow all the model licenses and limit the scope of these models to academic research only.

Data Sources	License
MSCOCO 2014	CC BY 4.0
Software Code	
LLAVA-NEXT	Llama 3 Community License
LLAVA-onevision	Apache License 2.0
Kimi-VL	MIT License
MiniCPM-o	Apache License 2.0
InternVL2.5	MIT License
mPLUG-Owl3	Apache License 2.0
Qwen2.5-VL	Apache License 2.0
GPT-4o	OpenAI Term of Use
Gemini	Google Term of Service

Table 1: License information for the scientific artifacts used.

B Experiment details

B.1 Dataset Construction Pipeline

Language Instruction Prompts.
Within our data construction pipeline, we design prompts for both Intervention Objects Selection and Counterfactual Description Generation stages. These prompts serve as the input for Gemini, and their specific content is provided below.

- Intervention Objects Selection stage:

> I now want to focus specifically on the **{target_object}** that is colored in red in the image, and replace only this specific **{target_object}** with another object. Here are the replacement items to choose from: **{candidate pool}**. Which item do you think would be more suitable in terms of size and how well it fits the image? Please output only one item and do not include any other output.

- Counterfactual Description Generation stage:

> Describe the picture based on the given objects:**{ground-truth object annotation}**, be as specific as possible, not too short.

> Now, I want to replace only the **{target_object}** that is specifically colored in red with the **{counterfactual_object}** using the inpainting model, while leaving any other instances of {target_object} in the image unchanged, could you give me a concrete and accurate description for the inpainted picture? Your description should include the **{counterfactual_object}** and the objects in the just-mentioned object list except for the {target_object}, be as specific as possible, not too short. And there is no need to mention the replacement operation in the description. You only need to give me the description without any additional content."

Pipeline Details.
In our pipeline, Gemini2.5-Flash [9] is employed for object selection and description generation. To extract the object masks, we leverage Grounded SAM 2 [8], which takes both textual prompts (the object's name) and an image as input. The extracted masks are then expanded using OpenCV's dilate function with a 5x5 kernel over ten iterative dilation steps. Finally, for image inpainting, we utilize the FLUX-ControlNet Inpainting Beta model [14, 4], setting controlnet_conditioning_scale to 0.9, guidance_scale to 3.5, and true_guidance_scale to 1.0.

B.2 Model Evaluation

We use official APIs to evaluate proprietary LVLMs, GPT-4o [3] and Gemini 1.5 pro [10]. For all the open-source models used in our experiments[7, 5, 11, 12, 2, 13, 1], we used their official source code and performed inference on each model using a single NVIDIA A40 GPU.

C Case Study

In this section, we conducted a case study on the Casual HalBench to investigate different types of questions, including contextual questions (Appendix C.1), absent questions (Appendix C.2), and counterfactual questions (Appendix C.3).

C.1 Contextual Questions

In this subsection, we present a case study related to contextual questions, as shown in Figures 1, 2, and 3.

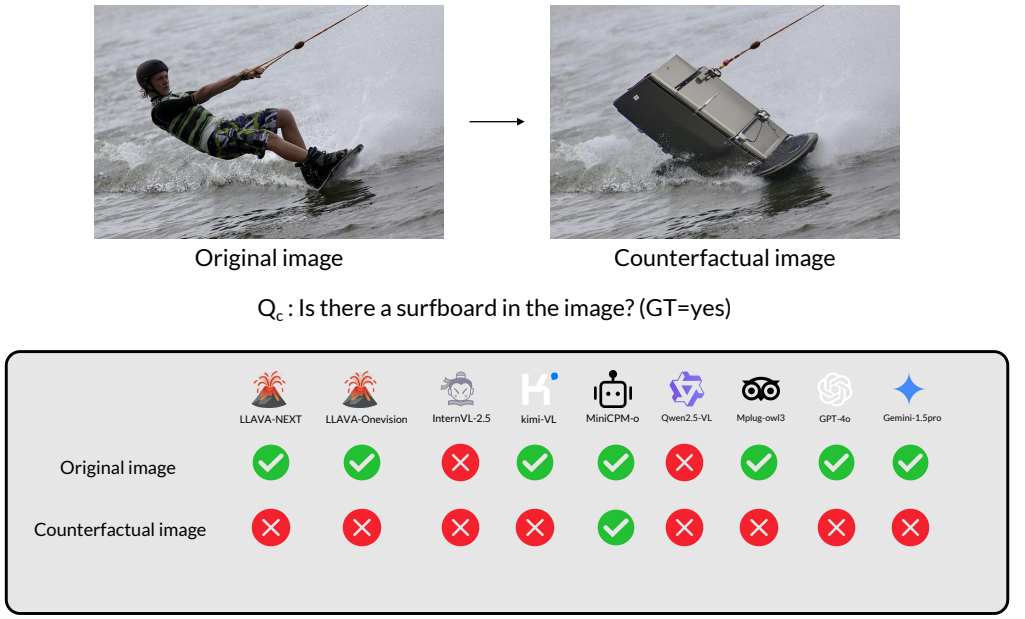
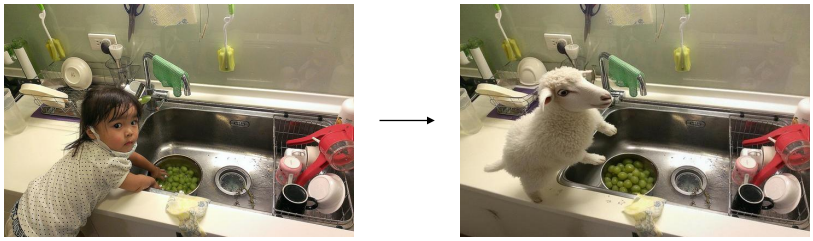


Figure 1: A case study of contextual questions on the Casual HalBench.



Original image Counterfactual image

Q_c : Is there a bottle in the image? (GT=yes)

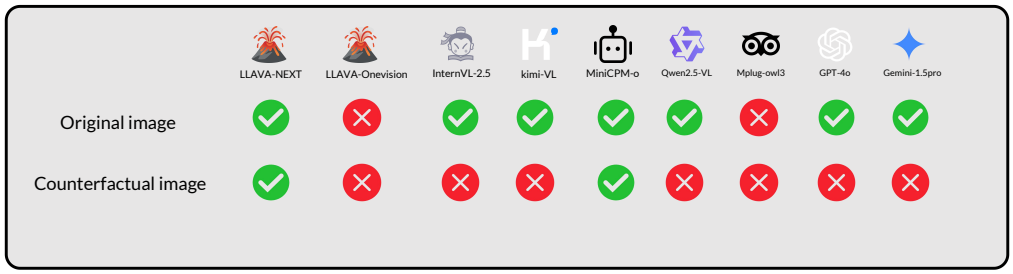
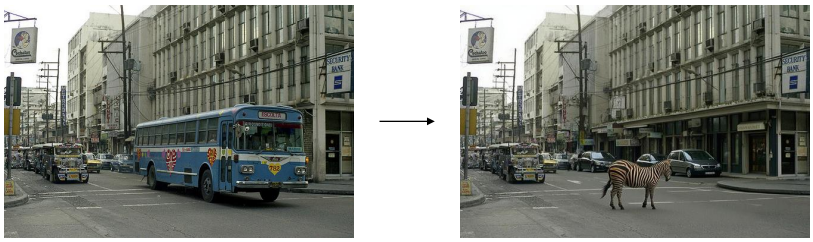


Figure 2: A case study of contextual questions on the Casual HalBench.



Original image Counterfactual image

Q_c : Is there a bottle in the image? (GT=yes)

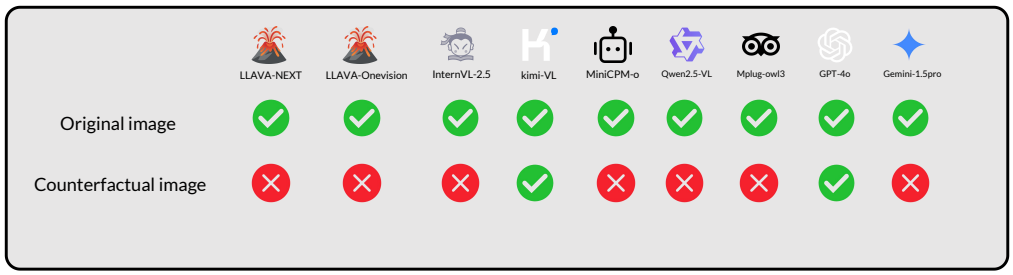
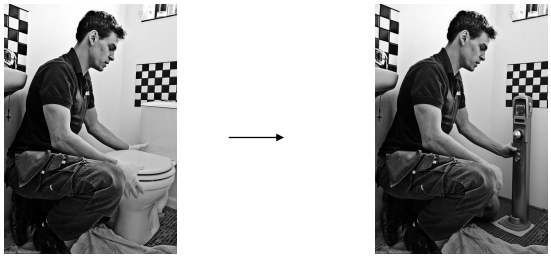


Figure 3: A case study of contextual questions on the Casual HalBench.

C.2 Absent Questions

In this subsection, we present a case study related to absent questions, as shown in Figures 4, 5, and 6.



Original image Counterfactual image

Q_c : Is there a bowl in the image? (GT=no)

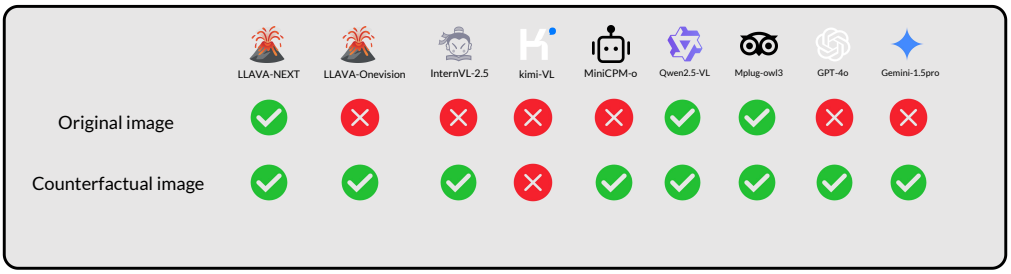
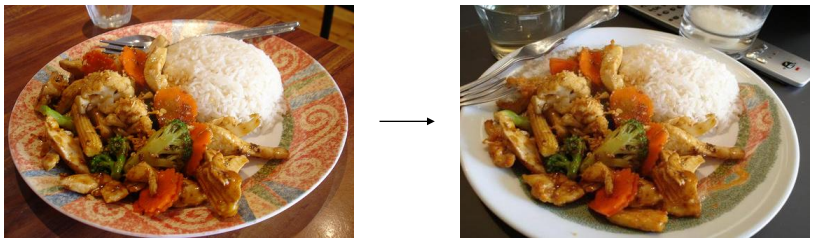


Figure 4: A case study of absent questions on the Casual HalBench.



Original image Counterfactual image

Q_c : Is there a chair in the image? (GT=no)

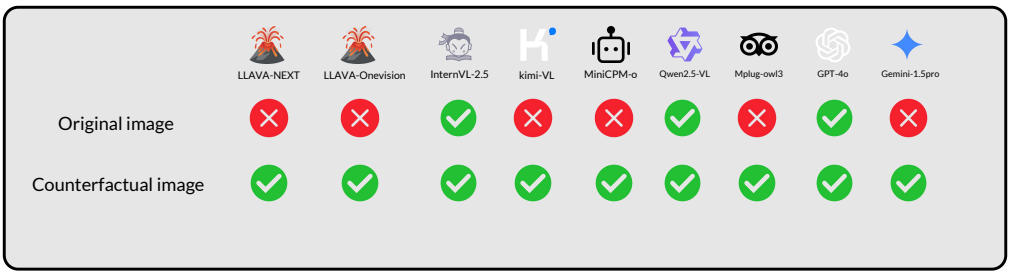
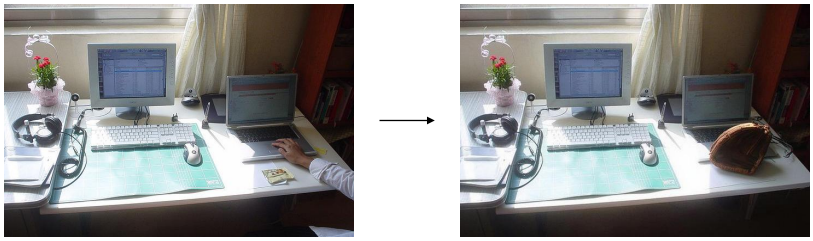


Figure 5: A case study of absent questions on the Casual HalBench.



Q_c : Is there a chair in the image? (GT=no)

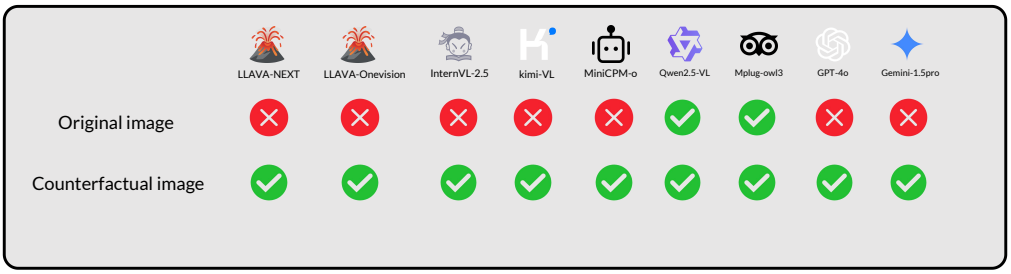
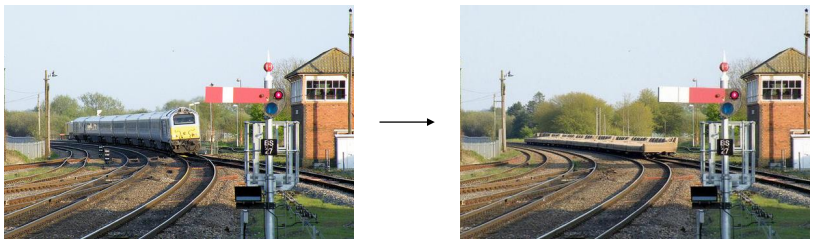


Figure 6: A case study of absent questions on the Casual HalBench.

C.3 Counterfactual Questions

In this subsection, we present a case study related to absent questions, as shown in Figures 7, 8, and 9.



Q_{cf} : Is there a couch in the image? (GT=yes)

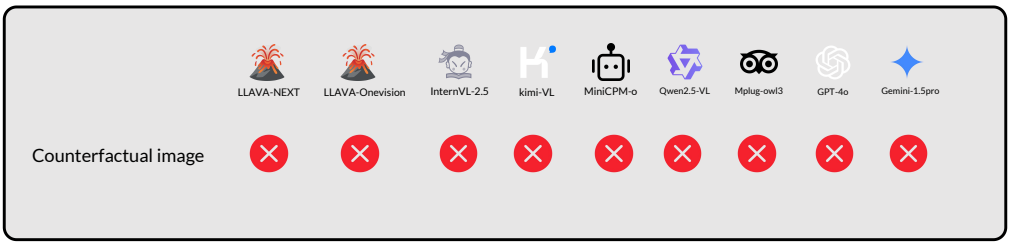
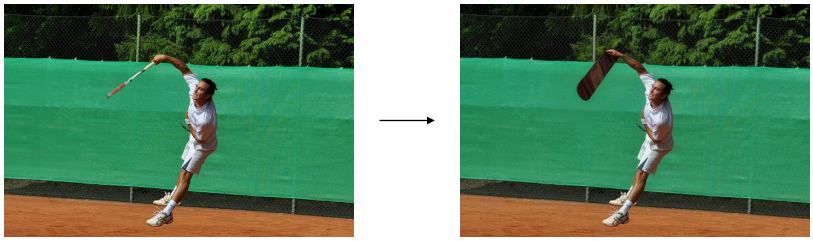


Figure 7: A case study of counterfactual questions on the Casual HalBench.



Original image Counterfactual image

Q_{cf} : Is there a snowboard in the image? (GT=yes)

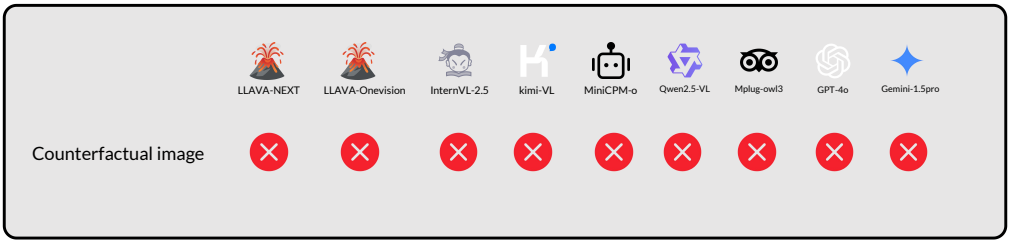


Figure 8: A case study of counterfactual questions on the Casual HalBench.

Original image Counterfactual image

Q_{cf} : Is there a cat in the image? (GT=yes)

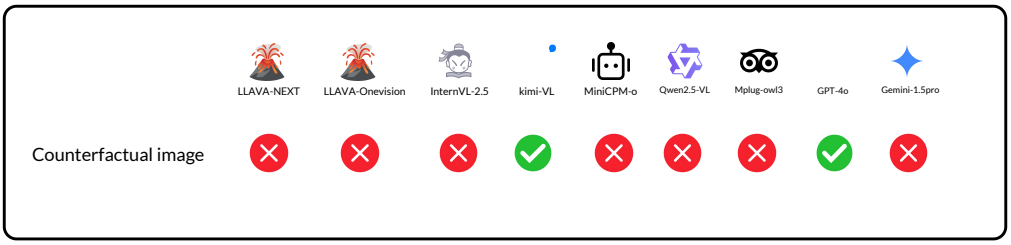


Figure 9: A case study of counterfactual questions on the Casual HalBench.

References

[1] Bai, S.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Song, S.; Dang, K.; Wang, P.; Wang, S.; Tang, J.; et al. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

[2] Chen, Z.; Wang, W.; Cao, Y.; Liu, Y.; Gao, Z.; Cui, E.; Zhu, J.; Ye, S.; Tian, H.; Liu, Z.; et al. 2024. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*.

[3] Hurst, A.; Lerer, A.; Goucher, A. P.; Perelman, A.; Ramesh, A.; Clark, A.; Ostrow, A.; Welihinda, A.; Hayes, A.; Radford, A.; et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

[4] Labs, B. F. 2024. FLUX. https://github.com/black-forest-labs/flux.

[5] Li, B.; Zhang, Y.; Guo, D.; Zhang, R.; Li, F.; Zhang, H.; Zhang, K.; Zhang, P.; Li, Y.; Liu, Z.; et al. 2024. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*.

[6] Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, 740–755. Springer.

[7] Liu, H.; Li, C.; Li, Y.; Li, B.; Zhang, Y.; Shen, S.; and Lee, Y. J. 2024. Llavanext: Improved reasoning, ocr, and world knowledge.

[8] Ren, T.; Liu, S.; Zeng, A.; Lin, J.; Li, K.; Cao, H.; Chen, J.; Huang, X.; Chen, Y.; Yan, F.; Zeng, Z.; Zhang, H.; Li, F.; Yang, J.; Li, H.; Jiang, Q.; and Zhang, L. 2024. Grounded SAM: Assembling Open-World Models for Diverse Visual Tasks. arXiv:2401.14159.

[9] Team, G.; Anil, R.; Borgeaud, S.; Alayrac, J.-B.; Yu, J.; Soricut, R.; Schalkwyk, J.; Dai, A. M.; Hauth, A.; Millican, K.; et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

[10] Team, G.; Georgiev, P.; Lei, V. I.; Burnell, R.; Bai, L.; Gulati, A.; Tanzer, G.; Vincent, D.; Pan, Z.; Wang, S.; et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

[11] Team, K.; Du, A.; Yin, B.; Xing, B.; Qu, B.; Wang, B.; Chen, C.; Zhang, C.; Du, C.; Wei, C.; et al. 2025. Kimi-vl technical report. *arXiv preprint arXiv:2504.07491*.

[12] Team, O. M.-o. 2025. Minicpm-o 2.6: A gpt-4o level mllm for vision, speech, and multimodal live streaming on your phone.

[13] Ye, J.; Xu, H.; Liu, H.; Hu, A.; Yan, M.; Qian, Q.; Zhang, J.; Huang, F.; and Zhou, J. 2024. mplug-owl3: Towards long image-sequence understanding in multi-modal large language models. *arXiv preprint arXiv:2408.04840*.

[14] Zhang, L.; Rao, A.; and Agrawala, M. 2023. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, 3836–3847.